Forecasting Movie Rating Through Data Analytics

Latika Kharb[1(✉)], Deepak Chahal[1], and Vagisha[2]

[1] Jagan Institute of Management Studies, Sector-5, Rohini, Delhi 110085, India
latika.kharb@jimsindia.org
[2] CSE, Banasthali Vidyapeeth, Newai, Rajasthan, India

Abstract. Movie prediction is an important way to predict movie revenue and performance. Through data analysis, we can find the most popular genres, performance in recent years and how it affects the reputation of the next movie. As movie production incurs huge cost and efforts, our effort is to predict the percentage of success, so that production could be managed accordingly. Throughout the paper, we will discuss about the different ways in which the data analysis used by the film gives a precise idea to each production about the best or worst chances of success and/or failure. In this paper, our goal is to focus on predicting the profitability of a film to support film investment decisions in the early stages of film production. The movie producers and directors can make use of the proposed model in various ways like: modify the movie criteria for becoming blockbusters, launch movie at particular time period to get maximum profit, predict the fan following to get a blockbuster and so on.

Keywords: Prediction · Data analysis · Blockbusters · Investment · IMDb

1 Introduction

Movie prediction is an important way to predict movie revenue and performance. Some of the criteria in calculating movie success included budget, actors, director, producer, set locations, story writer, movie release day, competing movie releases at the same time, music, release location and target audience [1]. IMDB is the most popular website for movie ratings and movie reviews. Imagine being able to analyze the reviews and understand what they liked or did not like the customers. By doing so, we can measure customer satisfaction or dissatisfaction with the movie, which can affect the revenue generated by the movie in a positive or negative way. In Fig. 1, the graph shows the genres of films by IMDB scores.

The analysis of movie data can be incredibly powerful and can make informed guesses, but cannot determine the fate of an individual project with absolute certainty. For some films, the success will be the sale of tickets, for others, the profit margin, the reviews, the social conversations, the franchise options or the Critical Awards.

U. Batra et al. (Eds.): REDSET 2019, CCIS 1230, pp. 249–257, 2020.
https://doi.org/10.1007/978-981-15-5830-6_21

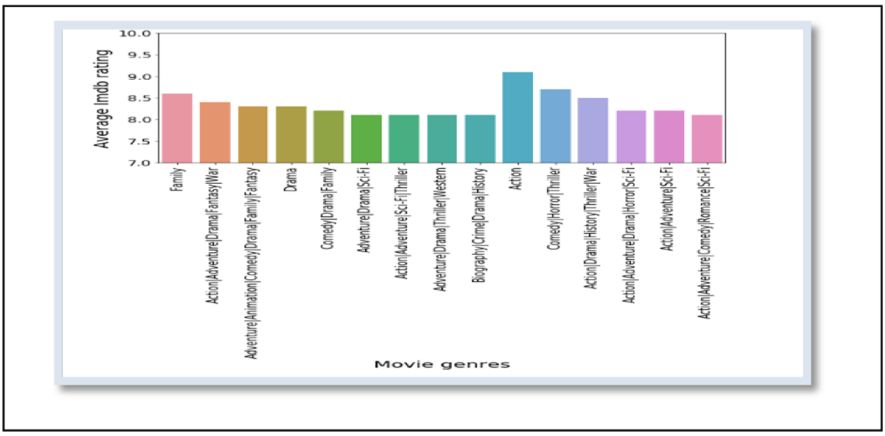

Fig. 1. Movies genres graph by IMDB score

2 Related Work

Since the analysis of movie data is such a hot topic in recent years, many articles have been published in the field of data analysis and its related field. The focus is to make the machines interdependent i.e. they don't require any kind of raw data sets to process the information [2]. In this section, we will discuss several relevant works that have been published. Prof. Junghare [3] suggested a model on the subject "Statistical analysis in reviews and ratings of films" to see how statistical analysis can be performed in reviews and ratings of films. The opinion of PEOPLE is one of the most important sources for different services. The statistical analysis of movie reviews and ratings gives users a perfect picture of what social media thinks about the movie. The movie rating information that will be generated is based on various sources such as Twitter, Facebook, IMDb and Google Trend. Federico de Gregorio [4] tackled the topic Predicting movie box office performance using, YouTube media and the database of IMDb movie data. The prediction model is primarily based on various decision key factors taken from the historical database of movies. The number of Twitter followers and the comparative analysis of comments from YouTube viewers. Postmus [5] proposed recommendation system techniques applied to Netflix movie data. This document contains the approach, the methodology, the elaboration and the evaluation of several common techniques of the recommendation system, applied to the Netflix qualifications. The data contains many user ratings on a Likert scale of 1 to 5 in different movies. The goal is to recommend movies to users who have not yet seen. SarathBabu PB [6] on the theme "Predicting the success of a movie based on the data of IMDb" points to a detailed study. Krauss [7] promoted the success of films and the awards of the academy through the analysis of feelings and social networks.

3 Background

We focus on predicting the profitability of a film to support film investment decisions in the early stages of film production. Using previous data from various sources, and using Python data analysis, this document extracts several types of features, including the theme of the movie, "when" a movie will be released, etc. [8] The results of the experiment showed that the system exceeds the reference methods by a wide margin. In addition to designing a decision support system with practical utility, this research highlights the power of predictive and prescriptive data analysis in information systems to help business decisions. For this document, we use data from Hollywood movies (2000–17). On this data, we simply perform some operations using Python to predict results. And with the help of these results, we can predict what kind of movies people liked. Artificial Intelligence involves Machine Learning and Deep Learning in which Machine learning is the subset of Artificial Intelligence, and Deep Learning is the subset of Machine Learning [9]. In this document, we chose python because python supports large libraries and data, for example, Numpy, Pandas, Scip, Matplotlib, scikit-learn, Seaborn, etc. For an effective analysis of the data.

4 Proposed Methodology

The project pipeline is organized as follows. To perform the data analysis, the data must be chosen or prepared to obtain a set of data. A number of authors tried to surface the issue but superficially, this paper seeks to rectify this omission [10].

4.1 Selecting and Importing Data

In this phase we collect information about movies and everything related. Basically, we gather all the information mainly from IMDB and part of local websites. After collecting information, we organize the data in the form of a CSV file.

We convert the data in CSV format because Python IDE, that is, the jupyter notebook supports CSV or .XLSX files. Now that the data is selected, organized and converted into a compatible format, the data is now ready to be imported into the Python IDE.

4.2 Cleaning

This first module manages the basic cleaning operations, which consist in eliminating unimportant or annoying elements for the following phases of analysis and in the normalization of some misspelled words (Fig. 2).

	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor_2_name	actor_1_facebook_likes	gross
0	James Cameron	723.0	178.0	0.0	855.0	Joel David Moore	1000.0	760505847.0
1	Gore Verbinski	302.0	169.0	563.0	1000.0	Orlando Bloom	40000.0	309404152.0
2	Sam Mendes	602.0	148.0	0.0	161.0	Rory Kinnear	11000.0	200074175.0
3	Christopher Nolan	813.0	164.0	22000.0	23000.0	Christian Bale	27000.0	448130642.0
5	Andrew Stanton	462.0	132.0	475.0	530.0	Samantha Morton	640.0	73058679.0
6	Sam Raimi	392.0	156.0	0.0	4000.0	James Franco	24000.0	336530303.0
7	Nathan Greno	324.0	100.0	15.0	284.0	Donna Murphy	799.0	200807262.0

Fig. 2. Cleaning the data

4.3 Selecting and Importing Libraries for Movie Analysis

After cleaning the data, we selected and imported some libraries that would help us generate and predict the results (Fig. 3).

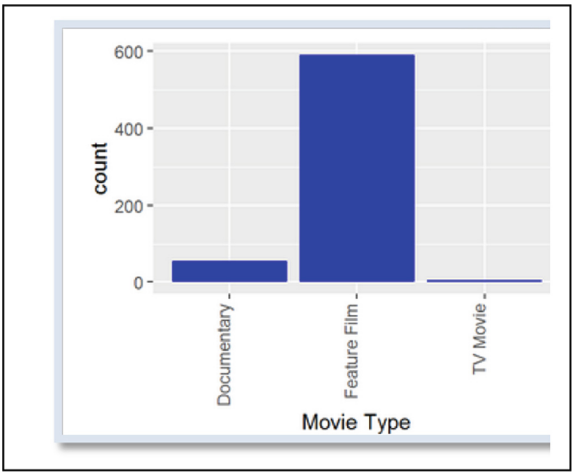

Fig. 3. Movie type selection

In Python, the Matplotlib.pyplot library helps us generate graphics [11, 12].

The panda library is used for data manipulation and analysis. In particular, it offers data structures and operations to manipulate numerical tables and time series (Fig. 4).

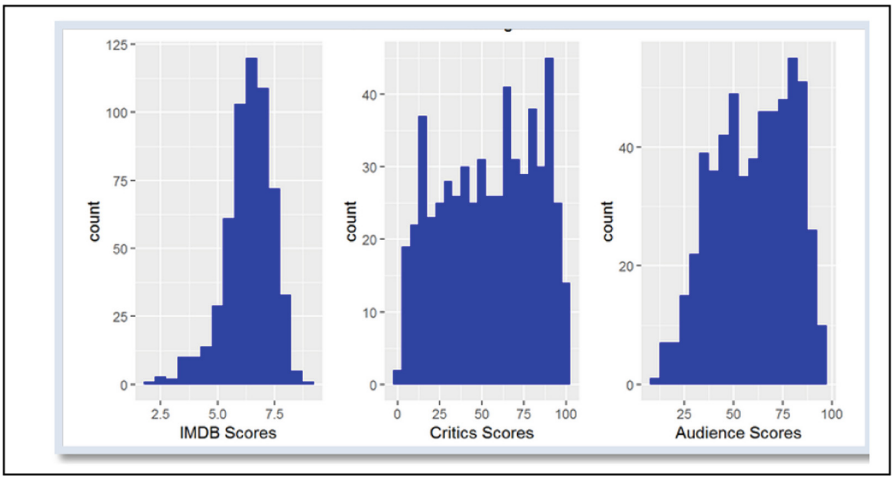

Fig. 4. Scores and count of movies

NumPy is the most basic but powerful package for scientific computing and data manipulation in Python.

The Seaborn library is used for data visualization and provides a high-level interface for drawing attractive and informative statistical graphs. The Seaborn library is based on matplotlib, the pyrotechnics library.

WordCloud is a data visualization technique used to represent text data in which the size of each word indicates its frequency or importance. You can highlight significant textual data points using a word cloud. Word clouds are widely used to analyze data from social networking websites. To generate a word cloud, we import the word cloud library in the Python IDE as shown in Fig. 5.

4.4 Implementation and Generating Results

After selecting and importing all libraries, we are ready to implement our algorithm to data for predicting the results. All the data has been sourced from secondary sources [13]. By using Matplotlib library we generated pie charts, line and bar graphs which shows us the various results, reaction of the people etc.

Seaborn library helped us to visualize the data and generate more effective graphs which helped us to analyze the data more effectively. By using we generated results as follow:

a) Which films are the best since 2000, as shown in Fig. 6. This result helped the directors and the people to choose which films are the best and direct the new version of the films, as they make a new version is a new trend in the city?

Fig. 5. WordCloud representation

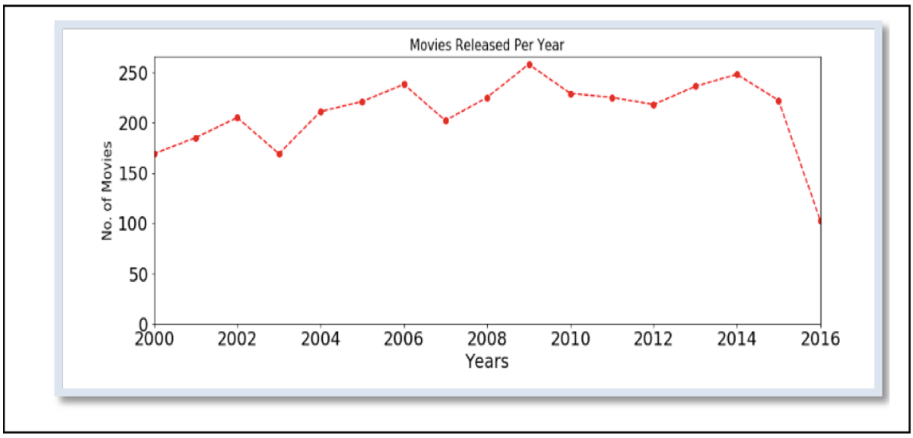

Fig. 6. Movie released per year

b) Find the best genres chosen by the viewers (according to the IMDB rating): This result helped the directors and producers to choose the best and most favorable genre so that the film can obtain a remarkable benefit
c) How many films are released per year: this result can be used by the government so that they can judge the amount of revenue generated and the amount of taxes that must be collected from the filmmakers.
d) Which language is popular among movie viewers and which country produces the most? of films: this result is used again by the producers and directors so that they can choose the language and make the film accordingly.

5 Experimental Result

In this paper, we have found the following results as follow:

a) Over the next analysis of the film we obtain the results and, according to me, the results were shocking. The film that we found the best since 2000 was "kickboxer: vengeance". According to IMDB, this movie received a rating of 9.0 out of 10.
b) According to the IMDB rating, the best genre is "Action" and your average IMDB score ranged between 9.0 and 9.5 out of 10. The second favorite genre is "thriller, comedy and horror".
c) According to the analysis, we also find that, on average, how many movies are released per year and in 2009 around 250 movies were released.
d) We also predicted the results that the language is popular among movie viewers and which country produces the maximum number of films. According to the results, we found that about 92.2% of people like English movies, 0.8% of people like Hindi movies and 7.0% of people like movies in other languages.

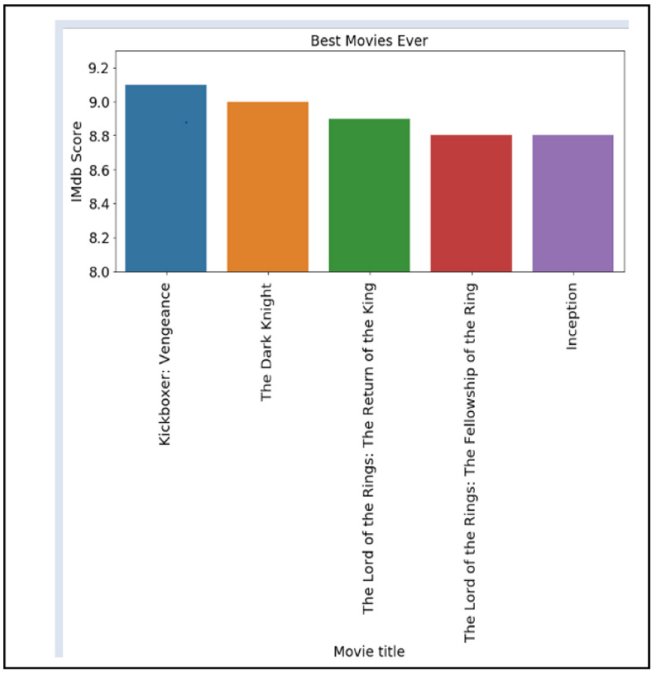

Fig. 7. Best movies

e) In addition, we discovered that countries produce more films than other countries and which country is more popular for their films, as shown in Fig. 7. We find that 74.0% like US movies. UU People love movies from other countries.

f) By film analysis we also discovered that in 2015 and 2016 how many genres and their films were produced during these years (Fig. 8).

Movie Title	Predicted Rating	95% Prediction Interval	Actual Rating
Dirty Grandpa	5.9	4.1 - 7.8	6.0
Deadpool	6.1	4.2 - 7.9	8.1

Fig. 8. Prediction of model

In this analysis, multi-variable, linear regression model was created that proved to have some capability for predicting movie popularity as indicated by IMDB movie rating score.

The prediction is rely on various decision making factors derived from historical database of movies, count of followers tweets from Twitter, and sentiment analysis of comments of YouTube viewers.

6 Conclusion and Future Work

In this article, the predictive model for the box office performance of films was represented by data derived from social networks and IMDb. According to our models, we identified the subsequent patterns: (a) the fame of the main artist is fundamental to the hit of a film, (b) the mixture of the booming past genre and a sequel movie is another guide for success, (c) a new movie in the less trendy genre and an artist with slight fame may perhaps be a sample for a failure.

Therefore, from the previous analysis, it is concluded that there is a need for a statistical analysis of the ratings and reviews of films. This would be a great and unique concept that will be introduced in the market. The results obtained after the implementation of our predictive model are better as compared to similar studies already done in this field. Although the results are not good enough for professional or business purposes, our model can be used in some online applications. A larger training set is required to enhance the performance of the model.

References

1. Ahmad, J., Duraisamy, P., Yousef, A., Buckles, B.: Movie success prediction using data mining. In: 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, pp. 1–4 (2017). https://doi.org/10.1109/icccnt.2017.8204173
2. Kharb, L., et al.: Brain emulation machine model for communication. Int. J. Sci. Technol. Res. (IJSTR), 1410–1418 (2019). http://www.ijstr.org/final-print/aug2019/Brain-Emulation-Machine-Model-For-Communication.pdf

3. Belgur, V., Karande, A., Junghare, A.M.: Statistical analysis on movie reviews and ratings. In: BE in Computer Engineering ZCOER, Pune, Maharashtra, India (2017)
4. Krushikanth, R., Jose, A.M., Supreme, M.: Prediction of movies box office performance using social media. Department of Computer Science, The University of Akron (2013)
5. Postmus, S.: Recommender system techniques applied to Netflixdata. Vrije Universiteit, Amsterdam (2018)
6. Nithin, V.R., Pranav, M.: Predicting movie success based on IMDb data. Department of CSE, National Institute of Technology Calicut (2014)
7. Krauss, J.S.: Predicting movie success and academy awards through sentiment and social network analysis. University of Cologne, Pohligstrasse 1, Cologne, Germany (2008)
8. Oghina, A.: European Conference on Information Retrieval. Springer, Heidelberg (2012)
9. Kharb, L., et al.: Implementing IoT and data analytics to overcome "vehicles danger. Int. J. Innov. Technol. Explor. Eng. **8**(11), 4298–4304 (2019)
10. Kharb, L., Sukic, E.: An agent based software approach towards building complex systems. tEM J. **4**(3), 287 (2015)
11. Pedregosa, F., Vanderplas, J.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
12. Millman, K.J.: Python for scientists and engineers. Comput. Sci. Eng. **13**(2), 9–12 (2011)
13. Kharb, L., et al.: A conjecture on the exchange rate of bitcoin. Int. J. Sci. Technol. Res. **8**(10) (2019, in press). ISSN 2277-8616